Clim. Past Discuss.,
https://doi.org/10.5194/cp-2019-80-RC3, 2019

Climate
of the Past
Discussions

Interactive comment on "The importance of input data quality and quantity in climate field reconstructions – results from a Kalman filter based paleodata assimilation method" *by* Jörg Franke et al.

Edward Cook (Referee)

drdendro@ldeo.columbia.edu

Received and published: 6 December 2019

To begin, let me declare that I am not an expert on the new data assimilation methods (DA) being used for climate field reconstruction (CFR) now. Therefore, I will not comment on the way in which the "state-of-the-art paleo data assimilation approach" has been applied in this paper. Rather, I will stick more so to what the title of the paper indicates, i.e. "the importance of input data quality and quantity in climate field reconstructions" as a generic problem that spans all methods of CFR. In so doing, I will point out what I regard as a problem with one of the main conclusions of this paper.

This study is based on three collections of tree-ring records: "(1) 54 of the best temperature sensitive tree-ring chronologies chosen by experts; (2) 415 temperature sensitive tree-ring records chosen less strictly by regional working groups and statistical screening; (3) 2287 tree-ring series that are not screened for climate sensitivity." These are the N-TREND, PAGES2K, and B14 data sets, respectively. I will not get into the issue of how the tree-ring series were processed (detrended and standardized) for temperature reconstruction other than to say that it is crucial to the recovery of multi-decadal to centennial timescale variability. This is possible from tree rings as numerous published studies have shown, but it is a difficult problem nonetheless. Regarding this study, the processing methods used are likely to vary considerably between the three data sets used, with the 54 best N-TREND tree-ring chronologies processed best by the experts, but the effects of these differences are not possible to determine in this paper. This is not a criticism. It is just the way it is given the data used.

The importance of input data quality and quantity in climate field reconstructions is at a basic level a given, so much of what this paper demonstrates is not terribly surprising. Thus, as a first-order conclusion, data quality and quantity do matter and more of both is better than less. However, as the authors show, quantity does not necessarily help if the quality of climate signal in the tree rings is not also considered given the target variable being reconstructed, in this case temperature. Thus, data screening for the signal of interest can have a big impact on the quality of the climate field reconstructions produced. The generic process of data screening in dendroclimatology goes back many years of course (e.g., Fritts, 1962), so again there is no surprise here. What is more controversial is the use of precipitation-sensitive tree-ring series to reconstruct past temperature through an inverse evapotranspiration demand mediated temperature signal rather than through a direct temperature effect on tree growth. I will not dwell on this here because it appears to work okay in certain cases, e.g. Trouet et al. (2013). However, there remains some concern about how the power spectrum of temperature reconstructions based on these quite different tree growth signals may differ. Let's just say that an inverse temperature signal is not as optimal as the direct one used in Wilson et al. (2016) and should be used with caution.

What has not been adequately considered in the paper, however, are differences in the size and location of the proxy domains used in the CFR experiments relative to the size and location of the climate field being reconstructed. For example, there is a great difference in size and location between the domain occupied by the 54 N-TREND series and the temperature domain being reconstructed. This basic issue was investigated by Kutzbach and Guetter (1980) in their classic paper on paleoenvironmental network design. It is not often cited today, yet should be mandatory reading for anyone who wishes to engage in CFR. In it, Kutzbach and Guetter (1980) show that reconstructing a large climate field from a much smaller proxy field is likely to be far less effective compared to the case where the proxy field is large and extends beyond the limits of the climate field being reconstructed. Such is clearly not the case regarding the N-TREND data used in this paper's CFR experiments.

The N-TREND data are exclusively from the 40°-75°N region rather than over the much larger domains of the other two tree-ring data sets. As such, those 54 tree-ring chronologies were never intended to be used in the way done in this paper because the temperature signals in many of the N-TREND series are comparatively local and therefore most reliable at that spatial scale of the overall N-TREND domain. See Anchukaitis et al. (2017) for Part 2 of the N-TREND study and the maps contained therein. Thus, the statement in the Abstract ''. . . nor the small expert selection [N-TREND] leads to the best possible climate field reconstruction" is really quite unfair because the experiments in this paper were set up in almost the worst possible way for N-TREND to succeed well. Thus, I find the results of this study difficult to interpret because of the vastly different spatial sampling that exists between N-TREND and the other two tree-ring datasets relative to the temperature field being reconstructed.

The authors also talk about assessments of reconstruction skill or skill improvement, but this is not really true in the classical sense where estimates are compared to actual data not used in the model calibration exercise. ÂăSo, there is no true out-of-sample

skill assessment made in their analyses and estimates of true reconstruction skill remain unknown. This is basically acknowledged by the authors in lines 127-128: "it must be noted that the final reconstruction is consistent only in the model world." Yet, true model validation tests could have been made by reserving a traditional validation interval for testing as is typically done in classical statistical CFR. This can be done in the context of data assimilation for CFR too as discussed in Steiger et al. (2018). The authors could, for example, calibrate the proxies only back to 1920 and check performance of the reconstructions over the withheld interval for skill and clues of overfitting. How ever done, some form of out-of-sample model validation testing should be mandatory when applying and testing any CFR method.

More specifically, a statistic called the root-mean-square-error skill score (RE) is used in this paper to compare the relative performances of the tree-ring data sets used in the DA experiments. But there is some unwanted and unnecessary confusion here. The 'true' RE (Reduction of Error) has a long history of use in both meteorology (Lorenz, 1956) and paleoclimatology (Fritts, 1976) as a measure of skill of 'out-of-sample' forecasts and hindcasts, respectively. To use the RE as classically defined requires an explicit calibration interval for model development and estimation of its mean state (climatology) and an explicit validation interval for testing the skill of the model estimates against withheld or 'out-of-sample' data. In this case, the minimum benchmark for model skill is RE > 0, i.e. skill > climatology. This does not appear to be the case here. Rather the authors seem to be using the model ensemble mean without proxy assimilation as the reference. ÂăAs such, there are not any explicitly defined calibration and validation intervals, and the authors are just assessing whether the simulations that assimilate the proxies do better than simulations that are merely forced with SSTs. Thus, the RE in this paper is very different from the classical RE of Lorenz (1956) and Fritts (1976) and should be called something else to avoid confusion.

Overall, I find this paper to be publishable after revisions that seriously consider the points raised in this review. However, I admit to not finding the results particularly insightful either. They are pretty much as one would expect given the tree-ring data sets and experimental design used in this study.

References

Anchukaitis, K., Wilson, R., Briffa, K. Buentgen, U., Cook, E., D'Arrigo, R., Davi, N., Esper, J., Frank, D., Gunnarson, B., Hegerl, G., Helema, S., Klesse, S., Krusic, P., Linderholm, H.W., Myglan, V., Osborn, T., Rydval, M., Schneider, L., Schurer, A., Wiles, G., Zhang, P. and Zorita, E. 2017. Last millennium Northern Hemisphere summer temperatures from tree rings: Part II, spatially resolved reconstructions. Quaternary Science Reviews 163:1-22.

Fritts, H.C. 1962. An approach to dendroclimatology: screening by means of multiple regression techniques. Journal of Geophysical Research 67(4):1413-1420.

Fritts, H.C. 1976. Tree Rings and Climate. Academic Press, London.

Kutzbach, J.E. and Guetter, P.J. 1980. On the design of paleoenvironmental data networks for estimating large-scale patterns of climate. Quaternary Research 14:169-187.

Lorenz, E.N. 1956. Empirical orthogonal functions and statistical weather prediction. Scientific Report No. 1, Department of Meteorology, MIT, Cambridge, Mass.

Steiger, N.J., Smerdon, J.E., Cook, E.R. and Cook, B.I. 2018. A reconstruction of global hydroclimate and dynamical variables over the Common Era. Scientific Data 5:180086 doi:10.1086/sdata.2018.86.

Trouet, V., Diaz, H.F., Wahl, E.R., Viau, A.E., Graham, R., Graham, N., and Cook, E.R. 2013. A 1500-year reconstruction of annual mean temperature for temperate North America on decadal-to-multidecadal time scales. Environmental Research Letters 8:2-10.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2019-80, 2019.